Published online 7 November 2011 *Nucleic Acids Research, 2012, Vol. 40, Database issue* ***D667–D674***
doi:10.1093/nar/gkr945

The *Candida* genome database incorporates multiple *Candida* species: multispecies search and analysis tools with curated gene and protein information for *Candida albicans* and *Candida glabrata*

Diane O. Inglis, Martha B. Arnaud*, Jonathan Binkley, Prachi Shah, Marek S. Skrzypek, Farrell Wymore, Gail Binkley, Stuart R. Miyasato, Matt Simison and Gavin Sherlock

Department of Genetics, Stanford University Medical School, Stanford, CA 94305-5120, USA

Received September 9, 2011; Accepted October 11, 2011

ABSTRACT

The *Candida* Genome Database (CGD, http://www.candidagenome.org/) is an internet-based resource that provides centralized access to genomic sequence data and manually curated functional information about genes and proteins of the fungal pathogen *Candida albicans* and other *Candida* species. As the scope of *Candida* research, and the number of sequenced strains and related species, has grown in recent years, the need for expanded genomic resources has also grown. To answer this need, CGD has expanded beyond storing data solely for *C. albicans*, now integrating data from multiple species. Herein we describe the incorporation of this multispecies information, which includes curated gene information and the reference sequence for *C. glabrata*, as well as orthology relationships that interconnect Locus Summary pages, allowing easy navigation between genes of *C. albicans* and *C. glabrata*. These orthology relationships are also used to predict GO annotations of their products. We have also added protein information pages that display domains, structural information and physico-chemical properties; bibliographic pages highlighting important topic areas in *Candida* biology; and a laboratory strain lineage page that describes the lineage of commonly used laboratory strains. All of these data are freely available at http://www.candidagenome.org/. We welcome feedback from the research community at candida-curator@lists.stanford.edu.

INTRODUCTION

Candida albicans is the most common fungal pathogen causing invasive and bloodstream infections in immuno-compromised patients, although in recent years, several non-albicans species and other yeasts have also emerged as major opportunistic pathogens (1,2). Studies in the US identify *Candida glabrata* as the second most common *Candida* species involved in invasive fungal infections. Moreover, antifungal drug resistance, especially to azoles, is common among *C. glabrata* clinical strains isolated from patients with prior azole treatment (1). The availability of genome sequences for these pathogenic fungi has made it possible to study genes that play a role in pathogenesis and drug resistance in *Candida* species, thereby increasing our understanding of the mechanisms of virulence in fungal pathogens.

The *Candida* Genome Database (CGD, http://www.candidagenome.org/) is an online resource for the scientific research community studying fungal molecular biology and pathogenesis. The primary mission of CGD is to facilitate and accelerate *Candida* research by providing both an extensively curated compendium of *Candida* gene, protein and sequence information, and easy-to-use web-based tools for accessing, analyzing and exploring these data.

When the CGD project began in 2004, our initial efforts focused on curation of *C. albicans,* because it is the best-characterized species of the group and has the largest corpus of gene-specific scientific literature.

*To whom correspondence should be addressed. Tel: +1 650 736 0075; Fax: +1 650 724 3701; Email: arnaudm@stanford.edu

We have now expanded the scope of the project to include other *Candida* species, and provide an extensive suite of tools and resources that have been redesigned to facilitate the analysis of multiple species concurrently. The CGD Locus Summary Page (LSP) has been updated with information about the identity of orthologous genes in *C. glabrata*, and with orthology-based functional predictions and gene descriptions. We currently display both manual and computational gene, protein and sequence information about *C. albicans* and the recently added species, *C. glabrata*. We also provide genomic and protein sequence downloads and BLAST (3) resources for multiple *Candida* species and strains, including *C. albicans* strains SC5314 (4) and WO-1 (5), C. dubliniensis (6), *C. guilliermondii* (5), *C. lusitaniae* (5), *C. parapsilosis* (5), *C. tropicalis* (5), *Debaryomyces hansenii* (7) and *Lodderomyces elongisporus* (5). We will be adding curated information for all these other *Candida* species in the future.

All of the data in CGD are freely available. We also have an extensive suite of online user documentation, and provide advice and user support by e-mail at candida-curator@lists.stanford.edu.

LITERATURE CURATION FOR MULTIPLE *CANDIDA* SPECIES

At CGD, PhD level curators perform ongoing manual curation of the scientific literature to collect, organize, summarize and present a comprehensive picture of each characterized gene. Manual curation includes the recording of gene names, addition and updates to our summary gene descriptions, capture of mutant phenotype data and the assignment of relevant GO annotations with evidence and citations.

The manual curation of the previously published literature pertaining to genes of *C. albicans* and *C. glabrata* is now complete (Table 1). We have combed the scientific literature for gene-specific information and gene bibliographies; Gene Ontology (GO) annotations describing the function, role and localization of gene products; and mutant phenotypes. These are now reported in CGD for all of the genes for which this information is available. At this time, there are 6203 predicted *C. albicans* protein-encoding genes localized to chromosomes in the current (Assembly 21) reference gene set, 22% with manually annotated gene and protein information. For *C. glabrata*, the reference annotation set contains 5212 predicted genes, each of which has a LSP (Figure 1), and 3% of which have manually curated annotations. CGD now includes a detailed Genome Snapshot for *C. glabrata* in addition to *C. albicans*, which provides a graphical and tabular summary of information about the total number of chromosomal features and feature types, changes to the reference sequence and a distribution of gene products by functional categories and cellular localization (Figure 2).

In addition, CGD curators have composed in-depth descriptive Locus Summaries for 272 selected *C. albicans* genes, which, in contrast to the very concise Locus Descriptions, are more detailed enumerations of the characteristics of each gene, presented in a bullet-point format on the CGD LSPs. They provide additional experimental details and gene regulatory information that cannot be accommodated within the space limits of the Locus Description line. These lists are displayed in the Locus Summary section located near the bottom of the page and are fully searchable through the CGD Text Search tool.

The curation of the entire body of scientific literature for these organisms is a large and ongoing endeavor as new papers are published, and we welcome suggestions from users as to papers that should be prioritized or other data that should be included. We greatly appreciate the beneficial interactions with members of the *Candida* research community who have already volunteered to review specific LSPs and provide feedback on the curation content for specific genes. The comments we have received have resulted in refinement of description lines, improvements to phenotype and GO annotations, and addition of new references that we had not encountered in our literature searches—improvements that benefit the entire community of CGD users.

TOOLS FOR SEARCH AND DISPLAY OF MULTISPECIES INFORMATION IN CGD

CGD was originally modeled after the *Saccharomyces* Genome Database (SGD) (8), a database that provides the *Saccharomyces cerevisiae* reference sequence with

Table 1. CGD curation statistics

	Candida albicans	Candida glabrata
Number of ORFs	6108	5212
Number of tRNAs	156	230
Verified ORFs	1403	178
Uncharacterized ORFs	4705	5034
Dubious ORFs	152	N/A
Manual GO annotations	4697	4689
Features with manual GO annotations	13 707	2622
Orthology-based GO annotations	13 246	19 655
Features with orthology-based GO annotations	3099	4157
Protein-domain (InterPro)-based GO annotations	6048	5087
Features with protein-domain (InterPro)-based GO annotations	2963	2583
Features with orthology-based description lines	1352	3982

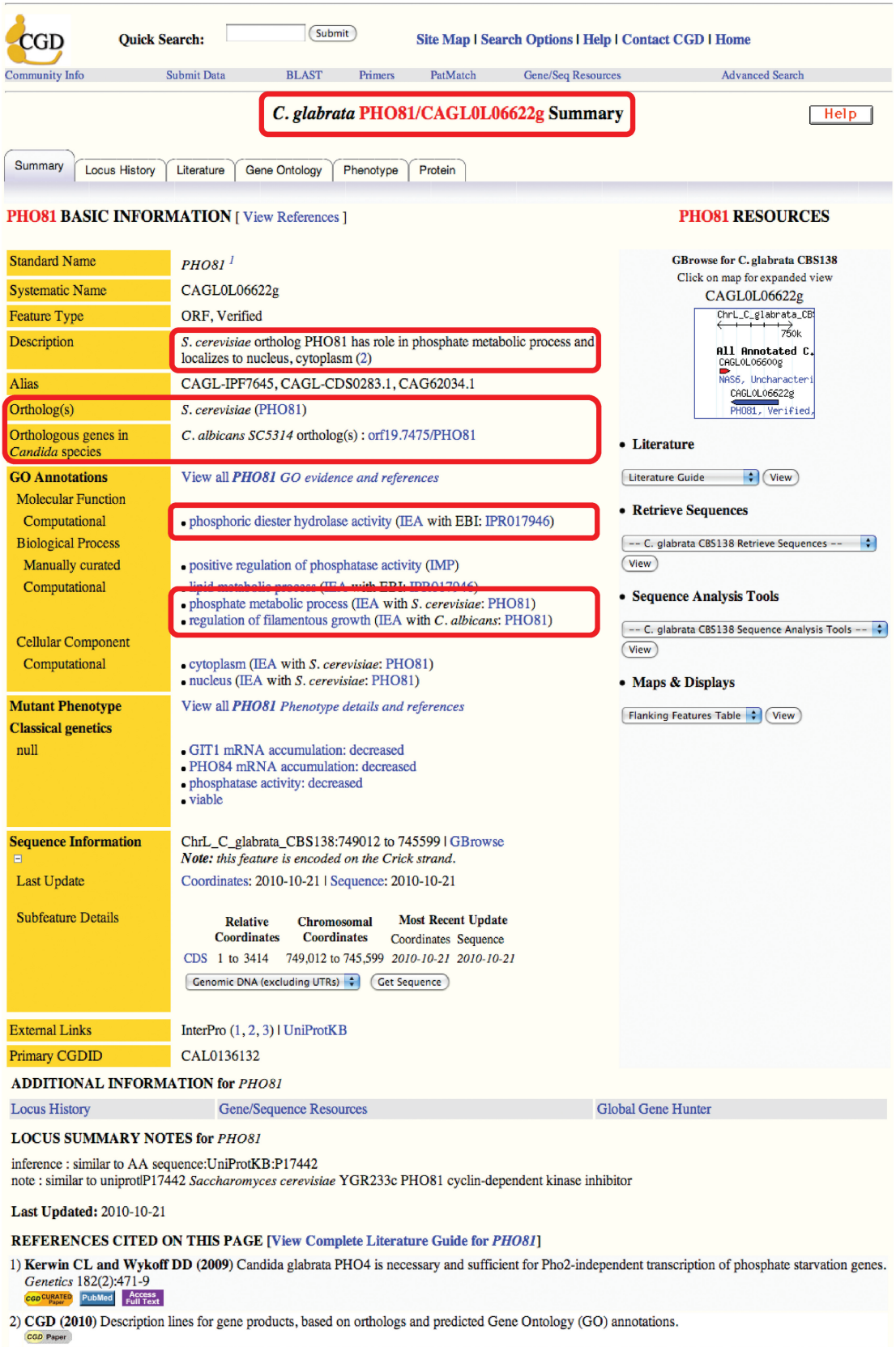


Figure 1. Updates to the CGD Locus Summary Page (LSP). The LSP is the hub around which the CGD gene information is organized. LSPs for both *C. albicans* and *C. glabrata* now feature new expanded orthology information sections, orthology-based description lines for uncharacterized genes, orthology-based GO term predictions and protein domain-based GO term predictions.

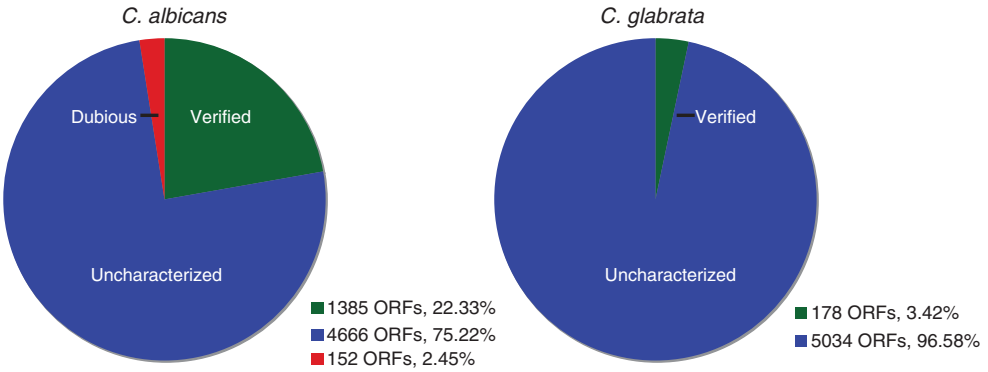


Figure 2. CGD genome snapshots. Pie chart from the CGD Genome Snapshots, comparing the current extent of the characterization of the predicted protein-coding genes in the *C. albicans* and *C. glabrata* genomes. ORFs are classified as 'Verified' if there is experimental evidence for a functional gene product. 'Uncharacterized' ORFs are predicted based on sequence analysis but currently lack experimental characterization. *Candida albicans* ORFs labeled as 'Dubious' have no experimental characterization and appear to be indistinguishable from random non-coding sequences (5).

literature curation, and gene, protein and sequence analysis tools for the *S. cerevisiae* research community. SGD, and initially CGD, were designed to store and display data for only a single species at a time. To accommodate the incorporation of additional species in the database, user interface and analysis tools, significant design modifications to the software and the underlying database structure were necessary.

The CGD search tools, such as Quick Search, Text Search, Gene/Sequence Resources, Ortholog Search and Pattern Match have been redesigned to search multiple species. In order to accommodate search results for multiple species, the new results page for the CGD Quick Search and Text Search tools now displays three sections. Search results that apply to all species (e.g. GO terms, authors and reference information, colleagues) are displayed at the top, with sections for species-specific search results displayed below. All of the tools that perform species- or sequence-specific searches (e.g. Gene/Sequence Resources, Pattern Match, Advanced Search, Batch Download, Restriction Mapper, GO Term Finder, GO Slim Mapper) have been updated, and they now prompt users to select the species of interest. The Ortholog Search now retrieves ortholog and best-hit matches among all of the species in CGD and SGD (currently *C. albicans*, *C. glabrata* and *S. cerevisiae*). BLAST searches at CGD have also been redesigned to allow queries against any combination of the several *Candida* species for which we have complete sequence sets (*C. albicans*, *C. glabrata*, *C. dubliniensis, C. guilliermondii, C. lusitaniae, C. parapsilosis, C. tropicalis, Debaryomyces hansenii* and *Lodderomyces elongisporus*). In addition, the curation tools have been extensively modified to facilitate the curation of multiple species.

Each gene in CGD is represented on a LSP, which is the central organizing unit of the CGD web site. The LSP contains the basic information that describes the gene and provides access to tools for retrieval, analysis and visualization of gene data. We have reengineered the LSPs to accommodate multispecies information (Figure 1). LSPs for each *C. albicans* and *C. glabrata* gene now feature an expanded orthology section, by which the LSPs of each *C. albicans* gene are hyperlinked to the LSPs of their *C. glabrata* orthologs, and vice versa. The LSP for C. glabrata genes also provide external links to gene pages available at Ge nolevures (http://www.genolevures.org/cagl.html#) This section also serves as a gateway to information about the orthologs in *Saccharomyces cerevisiae*, providing hyperlinks to the LSP of each ortholog in the SGD. Including *S. cerevisiae* ortholog information is especially useful for the *C. glabrata* LSPs: the evolutionary divergence between *C. glabrata* and *S. cerevisiae* is considerably more recent (100-300 million years ago) (7,9) than the divergence between these two species and *C. albicans* (700-800 million years ago) (10), and thus *C. glabrata* shares a larger number of orthologs with *S. cerevisiae* than with *C. albicans*, 4372 and 3201, respectively (as predicted by InParanoid). To define orthology relationships, we use the InParanoid algorithm, which identifies reciprocal best BLAST hits between species (11). These mappings and links are updated quarterly in order to reflect changes in gene models and annotations at CGD and SGD.

In addition to the new orthology relationships displayed in CGD, another level of similarity-based information is provided via the new Protein tab on the LSP of each protein-coding gene (Figure 3). This tab opens the Protein Information page that provides descriptions and a graphical display of conserved protein domains and motifs identified using InterProScan software (12,13). The Protein Information pages also display the structure of the most similar protein in the Protein Data Bank (14), and contain information about the predicted protein length, molecular weight, sequence and a link to a table of calculated physicochemical properties.

LEVERAGING MULTISPECIES INFORMATION IN CGD: HOMOLOGY-BASED FUNCTIONAL PREDICTIONS

The GO is a structured vocabulary that is used to describe three aspects of gene products: their molecular function or

C. albicans Grx1p/Orf19.6510p Protein Information

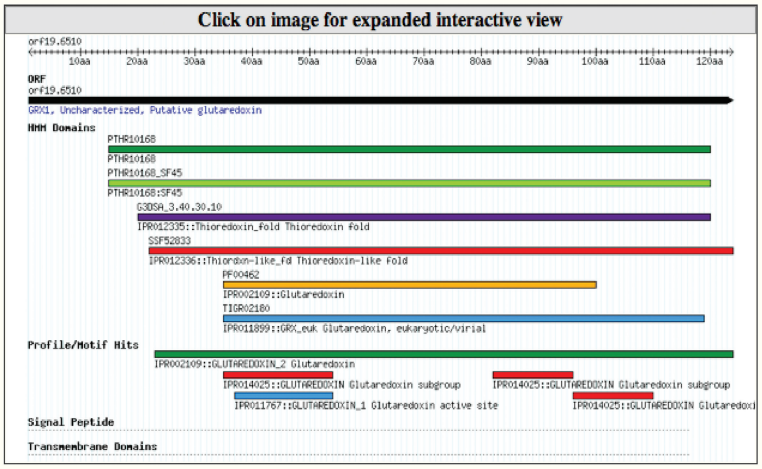
Help

Summary | Locus History | Literature | Gene Ontology | Phenotype | Protein

Grx1p PROTEIN INFORMATION [View References]

Standard Name	Grx1p [1]
Systematic Name	Orf19.6510p
Allele Name	Orf19.13863p
Description	Putative glutaredoxin; transcriptionally regulated by iron; expression greater in low iron (1)
Structural Information	View all PDB BLAST hits
Top PDB Hit: **3C1R**	View Top PDB Hit Information (Link-out)
Source	*Saccharomyces cerevisiae*; Crystal Structure Of Oxidized Grx1
Quality	41% identical to Grx1p; P-value = 1.60e-17 (View Alignment)
Structure	Click on image to access PDB interactive viewer (Link-out)

Conserved Domains	Domains/Motifs Page



Sequence Detail	Length = 123 aa; MW = 14.2 kDa; Physicochemical Properties Page
Download in FASTA format)	1 MSSILAWGFN LWYQPPPPTA QTEKEIEHTI NSHKIVIYSK TYCPFCDQTK 51 HLLNEQYPQE SYEVINLNIL DDGLTIQNQL YANTGQYMVP IIFINGQHVG 101 GNSEVQQLHT NGKLQELLNP QKY*
Homologs	BLAST Grx1p against other *Candida* sequences
External Sequence Database	UniProt: Q5AH28

REFERENCES CITED ON THIS PAGE [View Complete Literature Guide for Grx1p]

1) **Lan CY, *et al.* (2004)** Regulatory networks affected by iron availability in Candida albicans. *Mol Microbiol* 53(5):1451-69

CGD CURATED Paper | PubMed | Access Full Text | Web Supplement

Figure 3. Protein information page. The Protein Information page provides data including structural information inferred from homologs in PDB (RCSB Protein Data Bank), an interactive domains/motifs browser, protein sequence and physicochemical property details, BLASTP search against other CGD sequences and links to external protein resources such as UniProt.

activity, the broader biological process in which they participate, and the cellular location in which they reside (15). A gene product can be annotated with any number of terms about any of the three aspects, depending on the available data. Each GO term assignment is associated with an evidence code that describes the type of data the assignment is based on, and with a reference to its source. The GO is in wide use in genomic research and because it is rigorously structured, it ensures consistency in capturing of functional information about genes from different organisms and thus enables reliable analysis of biological significance of genomic data (15–21).

For the fully curated species, *C. albicans* and *C. glabrata,* all of the available gene-related literature pertaining to these two species has been read and all possible GO assignments from these papers have been made. To augment the manual curation, we have leveraged the orthology relationships to infer GO annotations for genes having an experimentally characterized ortholog in SGD or CGD. Predictions for *C. albicans* are made based on *S. cerevisiae* and *C. glabrata* orthologs, whereas predictions for *C. glabrata* are based on orthologs from *S. cerevisiae* and *C. albicans*. Despite the evolutionary distances between *C. albicans*, *C. glabrata* and *S. cerevisiae*, the use of orthology relationships to infer GO annotations between *C. albicans* and *C. glabrata* allow the transfer of a significant number of important pathogenesis-related terms to be transferred between these two fungal pathogens. Candidate GO annotations to be used as the basis for these inferences are limited to those with experimental evidence, i.e. associated with evidence codes of 'Inferred from Direct Assay (IDA)', 'Inferred from Physical Interaction (IPI)', 'Inferred from Genetic Interaction (IGI)', or 'Inferred from Mutant Phenotype (IMP)'. Any annotations that are themselves predicted in *S. cerevisiae* or in *Candida*, either based on sequence similarity or by some other methods, are excluded from this group to avoid transitive propagation of predictions. Also excluded from the predicted annotation set are annotations that are redundant with existing, manually curated annotations, or those that assign a related but less specific GO term other than candidate annotations. These orthology-based GO assignments are associated with evidence code 'Inferred from Electronic Annotation (IEA)' and displayed with the source species and gene name they are derived from along with a hyperlink to the appropriate LSP at CGD or SGD.

CGD has also taken advantage of protein domain and motif homology to assign GO annotations for *C. albicans* and *C. glabrata* genes. We systematically predict conserved domains in CGD protein sequences using InterProScan (12), and then use the InterPro-to-GO mappings (12,13) provided by the GO Consortium to provide molecular function annotations for those proteins. These annotations are assigned the evidence code IEA and are displayed with the InterPro identifier of the protein that serves as the basis for the annotation. The identifier is linked to the EMBL-EBI database to provide access to more extensive information about each structural domain. We have also used the tRNAscan-SE software to predict tRNA genes, and have inferred predicted GO annotations for these tRNAs (22).

The new annotations that have been transferred from *S. cerevisiae* to *C. albicans* and *C. glabrata*, and between *C. albicans* and *C. glabrata*, are summarized in Table 1. In addition to having the evidence code IEA, all these orthology-based annotations are identified as being derived computationally, rather than manually extracted from the scientific literature. Predictions are updated several times a year to make sure they remain current with annotation updates and new curation in CGD, SGD and in the protein domain datasets.

Now that all literature-based GO assignments for *C. albicans* and *C. glabrata,* and all orthology-based and protein domain-based predictions have been made, we consider curation of both species to be 'GO-complete'. For the remaining uncharacterized genes, we have explicitly assigned 'unknown' annotations to indicate that to the best of our knowledge no data are available.

We have also used the multispecies information to create informative descriptions for those *Candida* genes that lack any experimental characterization, and which therefore have no literature-based description on the LSP, incorporating orthology relationships and orthology-based functional predictions into the gene description in cases where there would otherwise be no information available.

CURATED INFORMATIONAL PAGES AT CGD

Additional CGD resources for the *Candida* research community include a new collection of bibliographies on topics relevant to *Candida* biology, which is accessible under 'Community Resources' from the navigation sidebar on the CGD Home page. These Highlights in *Candida* Biology contain lists of important references, including many key reviews, and are designed to provide an overview of selected subject areas in *C. albicans* and *C. glabrata* biology. This resource will be particularly valuable for those new to *Candida* research. As new species are curated at CGD, Highlights in *Candida* Biology will expand to include bibliographies on these species as well. The curated bibliographies are available at http://candidagenome.org/TopicBiblios.shtml.

We have also curated a directory of strains, which provides descriptions and references for commonly used *Candida* laboratory strains, along with a lineage diagram that graphically depicts the relationship among these strains. This information is available on the CGD web site at http://candidagenome.org/Strains.shtml. This resource is especially important for researchers because differences in strain background are known to have a significant impact on observed mutant phenotypes. In some cases, genes have been found to be lethal in one genetic background while successful gene disruption is possible in another. An example of this is the *C. albicans UME6* gene, for which homozygous mutants are viable in the SN152 genetic background (23) yet inviable in the BWP17 strain background (24). Because of its importance, we also

provide all available strain background information along with all of the curated phenotypes for each gene.

FUTURE DIRECTIONS

Now that the underlying database has been re-tooled to accommodate the curation of multiple species, we will add curated information for other *Candida*-related species including *C. dubliniensis, C. guilliermondii, C. lusitaniae, C. parapsilosis, C. tropicalis, Debaryomyces hansenii and Lodderomyces elongisporus*. In order to facilitate navigation across multiple genomes, we will provide links to an interactive comparative visualization tool, which will allow users to explore ortholog clusters in their genomic context.

Recent advances in genomics technologies have created a deluge of information that poses a significant challenge of making all these data organized and readily available to researchers. We have adapted our genome browser, GBrowse, to enable users to visualize unannotated transcripts in *C. albicans* that have been identified by RNAseq (25–27). These transcripts are aligned to the reference genome and displayed alongside the existing set of features in the reference annotation. We will further develop and/or integrate existing software to incorporate and visualize more types of data and more data sets from high-throughput studies.

ACKNOWLEDGEMENTS

The authors would like to thank Génolevures for making the *C. glabrata* CBS138 sequence available, Brendan Cormack and Suzanne Noble for strain lineage information, and Mike Cherry and SGD for their help. CGD is grateful to the many members of the *Candida* research community who have generously provided their feedback and support for the project.

FUNDING

Funding for open access charge: National Institute of Dental and Craniofacial Research at the US National Institutes of Health (grant no. R01 DE015873).

Conflict of interest statement. None declared.

REFERENCES

1. Ruhnke,M. (2006) Epidemiology of *Candida albicans* infections and role of non-Candida-albicans yeasts. *Curr. Drug Targets*, **7**, 495–504.
2. Miceli,M.H., Diaz,J.A. and Lee,S.A. (2011) Emerging opportunistic yeast infections. *Lancet Infect. Dis.*, **11**, 142–151.
3. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
4. Jones,T., Federspiel,N.A., Chibana,H., Dungan,J., Kalman,S., Magee,B.B., Newport,G., Thorstenson,Y.R., Agabian,N., Magee,P.T. *et al.* (2004) The diploid genome sequence of *Candida albicans*. *Proc. Natl. Acad. Sci., U S A.* **101**, 7329–7334.
5. Butler,G., Rasmussen,M.D., Lin,M.F., Santos,M.A., Sakthikumar,S., Munro,C.A., Rheinbay,E., Grabherr,M., Forche,A., Reedy,J.L. *et al.* (2009) Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes. *Nature*, **459**, 657–662.
6. Jackson,A.P., Gamble,J.A., Yeomans,T., Moran,G.P., Saunders,D., Harris,D., Aslett,M., Barrell,J.F., Butler,G., Citiulo,F. *et al.* (2009) Comparative genomics of the fungal pathogens *Candida dubliniensis* and *Candida albicans*. *Genome Res.*, **19**, 2231–2244.
7. Dujon,B., Sherman,D., Fischer,G., Durrens,P., Casaregola,S., Lafontaine,I., De Montigny,J., Marck,C., Neuvéglise,C., Talla,E. *et al.* (2004) Genome evolution in yeasts. *Nature*, **430**, 35–44.
8. Engel,S.R., Balakrishnan,R., Binkley,G., Christie,K.R., Costanzo,M.C., Dwight,S.S., Fisk,D.G., Hirschman,J.E., Hitz,B.C., Hong,E.L. *et al.* (2010) Saccharomyces Genome Database provides mutant phenotype data. *Nucleic Acids Res.*, **38**, D433–D436.
9. Wolfe,K.H. and Shields,D.C. (1997) Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, **387**, 708–713.
10. Hedges,S.B., Blair,J.E., Venturi,M.L. and Shoe,J.L. (2004) A molecular timescale of eukaryote evolution and the rise of complex multicellular life. *BMC Evol. Biol.*, **4**, 2.
11. Remm,M., Storm,C.E. and Sonnhammer,E.L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.*, **314**, 1041–1052.
12. Zdobnov,E.M. and Apweiler,R. (2001) InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847–848.
13. Hunter,S., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Das,U., Daugherty,L., Duquenne,L. *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res.*, **37**, D211–D215.
14. Rose,P.W., Beran,B., Bi,C., Bluhm,W.F., Dimitropoulos,D., Goodsell,D.S., Prlic,A., Quesada,M., Quinn,G.B., Westbrook,J.D. *et al.* (2010) The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res.*, **39**, D392–D401.
15. Gene Ontology Consortium. (2001) Creating the gene ontology resource: design and implementation. *Genome Res.*, **11**, 1425–1433.
16. Aslett,M. and Wood,V. (2006) Gene Ontology annotation status of the fission yeast genome: preliminary coverage approaches 100%. *Yeast*, **23**, 913–919.
17. Bult,C.J., Eppig,J.T., Kadin,J.A., Richardson,J.E. and Blake,J.A. (2008) The Mouse Genome Database (MGD): mouse biology and model systems. *Nucleic Acids Res.*, **36**, D724–D728.
18. Hong,E.L., Balakrishnan,R., Dong,Q., Christie,K.R., Park,J., Binkley,G., Costanzo,M.C., Dwight,S.S., Engel,S.R., Fisk,D.G. *et al.* (2008) Gene Ontology annotations at SGD: new data sources and annotation methods. *Nucleic Acids Res.*, **36**, D577–D581.
19. Sprague,J., Bayraktaroglu,L., Bradford,Y., Conlin,T., Dunn,N., Fashena,D., Frazer,K., Haendel,M., Howe,D.G., Knight,J. *et al.* (2008) The Zebrafish Information Network: the zebrafish model organism database provides expanded support for genotypes and phenotypes. *Nucleic Acids Res.*, **36**, D768–D772.
20. Tweedie,S., Ashburner,M., Falls,K., Leyland,P., McQuilton,P., Marygold,S., Millburn,G., Osumi-Sutherland,D., Schroeder,A., Seal,R. *et al.* (2009) FlyBase: enhancing *Drosophila* Gene Ontology annotations. *Nucleic Acids Res.*, **37**, D555–D559.
21. Phillips,G.N. Jr, Fox,B.G., Markley,J.L., Volkman,B.F., Bae,E., Bitto,E., Bingman,C.A., Frederick,R.O., McCoy,J.G., Lytle,B.L. *et al.* (2007) Structures of proteins of biomedical interest from the Center for Eukaryotic Structural Genomics. *J. Struct. Funct. Genomics*, **8**, 73–84.
22. Lowe,T.M. and Eddy,S. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **5**, 955–964.
23. Banerjee,M., Thompson,D.S., Lazzell,A., Carlisle,P.L., Pierce,C., Monteagudo,C., Lopez-Ribot,J.L. and Kadosh,D. (2008) *UME6*, a novel filament-specific regulator of *Candida albicans* hyphal extension and virulence. *Mol. Biol. Cell.*, **19**, 1354–1365.
24. Nobile,C.J. and Mitchell,A.P. (2005) Regulation of cell-surface genes and biofilm formation by the *C. albicans* transcription factor Bcr1p. *Curr. Biol.*, **15**, 1150–1155.

25. Mitrovich,Q.M., Tuch,B.B., De La Vega,F.M., Guthrie,C. and Johnson,A.D. (2010) Evolution of yeast noncoding RNAs reveals an alternative mechanism for widespread intron loss. *Science*, **330**, 838–841.
26. Sellam,A., Hogues,H., Askew,C., Tebbji,F., van Het Hoog,M., Lavoie,H., Kumamoto,C.A., Whiteway,M. and Nantel,A. (2010) Experimental annotation of the human pathogen *Candida albicans* coding and noncoding transcribed regions using high-resolution tiling arrays. *Genome Biol.*, **11**, R71.
27. Bruno,V.M., Wang,Z., Marjani,S.L., Euskirchen,G.M., Martin,J., Sherlock,G. and Snyder,M. (2010) Comprehensive annotation of the transcriptome of the human fungal pathogen *Candida albicans* using RNA-seq. *Genome Res.*, **20**, 1451–1458.